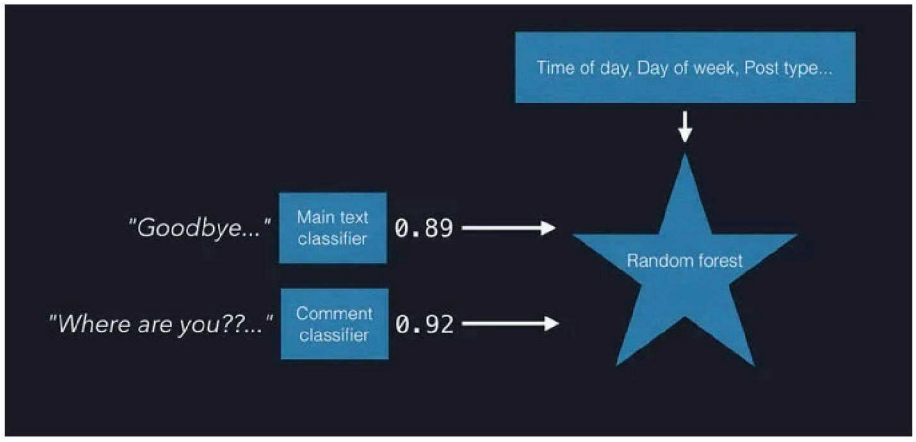
Tennessee v. Meta
Case No. 23-1364-IV

DX1093

POSTED ON FEBRUARY 21, 2018 TO ML APPLICATIONS

Under the hood: Suicide prevention tools powered by AI



By Dan Muriello, Lizzy Donahue, Danny Ben-David, Umut Ozertem, Reshef Shilon

Update on September 29, 2021: Information in this article may be outdated. For current information about our suicide and self-injury content detection technology, please visit our Safety Center. As described in the Safety Center, our algorithms are intended to help identify potential suicide and self-injury content and are not intended to diagnose or treat any mental health or other condition.

Originally published on February 21, 2018:

Suicide is the second most common cause of death for people ages 15-29. Research has found that one of the best ways to prevent suicide is for those in distress to hear from people who care about them. Facebook is well positioned — through friendships on the site — to help connect a person in distress with people who can support them. It's part of our ongoing effort to help build a safe community on and off Facebook.

We recently announced an expansion of our existing suicide prevention tools that use artificial intelligence to identify posts with language expressing suicidal thoughts. We'd like to share more details about this, as we know that there is growing interest in Facebook's use of AI and in the nuances associated with working in such a sensitive space.

In the past, we've relied on loved ones to report concerning posts to us, since they are in the best position to know when someone is struggling. However, many posts expressing suicidal thoughts are never reported to Facebook, or are not reported fast enough. We saw this as an opportunity to use machine learning to better detect posts indicating that someone may be at risk. When the system finds a post and a trained member of our Community Operations team determines that the poster might be at risk, we show that person resources with support options the next time they open Facebook.

To start, we worked with experts to identify specific keywords or phrases known to be associated with suicide. However, we found that we needed to do more than rely on simple rules, particularly because people often use the same words in different contexts (i.e., "Ugh, I have so much homework I just wanna kill myself"). As an alternative, we moved toward a machine learning approach, which excels at aggregating a series of weak signals to classify content. (You can read more about how we use this kind of artificial intelligence at Facebook here.) We were able to use posts previously reported to Facebook by friends and family, along with the decisions made by our trained reviewers (based on our Community Standards), as our

training data set. We used FBLearner, our internally developed machine learning engine, to train a classifier to recognize posts that include keywords or phrases indicating thoughts of self-harm.

We rely on human review by specially trained members of our Community Operations team to make the determination that a post expresses suicidal thoughts. It's important to minimize the number of false-positive reports sent to the team, so they can get help to people as quickly as possible. The best way to do this is by adding additional features to the model to get a stronger signal. Aggregated, de-identified comments on posts provide a crucial layer of context: For example, comments like, "Are you OK?" or, "Can I help you?" are a strong indicator that friends and family are taking a potential threat seriously. Additional research shows that both the time of day and the day of the week are important predictors of when someone is in crisis, so we include these variables in the model.

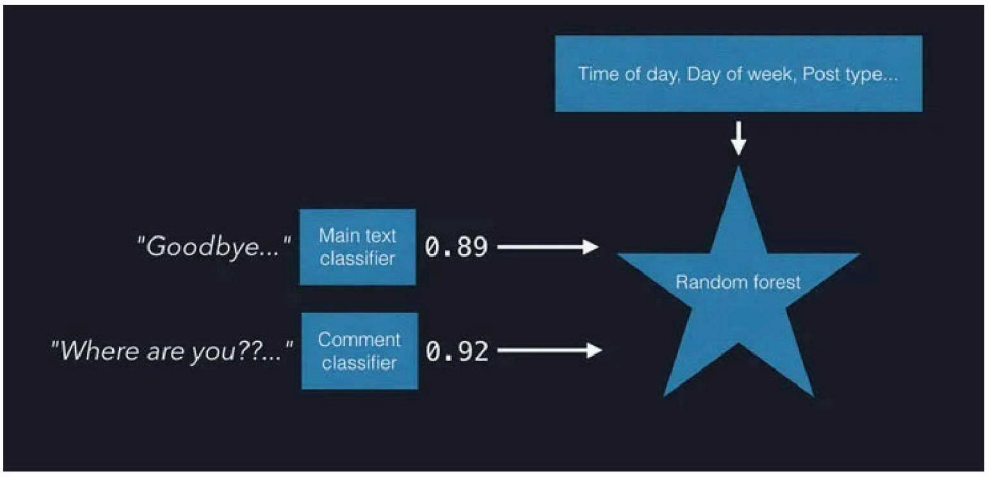
In our first attempt at text classification, we used a learning algorithm known as n-gram-based linear regression. This algorithm divides a piece of text into chunks of varying length, then looks to see how often each chunk appears in the positive training samples (that is, posts that express suicidal thoughts) versus the negative training samples (posts that do not express suicidal thoughts). Chunks appearing much more often in the positive samples are given a higher weight, while chunks that appear much more often in the negative samples are given a negative weight.

Here's a simple visualization of this process based on a made-up sentence and simulated weights:



The resulting text classification model can apply these weights to a new block of text and create a score indicating its similarity to text contained in the positive training samples (posts indicating suicidal thoughts).

We created two text classifiers: one predicting the main text of posts, and another the comments. The scores from each of these classifiers, combined with other numerical features (such as the time of day, day of the week, or post type; for instance whether the post was posted on a person's own timeline or a friend's timeline), provide the inputs to a random forest learning algorithm, which specializes in learning using numerical features.



The random forest algorithm builds a classifier that scores posts according to their similarity with prior posts previously identified as expressing suicidal thoughts. If this classifier gives a post a sufficiently high score, it is sent to our Community Operations team for review.

We updated this system to use a DeepText-based neural network. This creates a vector space of word meanings and uses neural networks to find the relevant patterns in those sentences. We incorporated three additional DeepText classifiers into our subsequent random forest classifiers: two that reexamine the post and comment texts respectively similar to the n-gram-based models, and a third that is trained on individual comments from posts escalated to secondary review for possible in-person intervention. We did this because people in imminent danger tended to get comments such as "Tell me where you are" or "Has anyone heard from him/her?" while less urgent posts had comments more along the lines of "Call anytime" or "I'm here for you." We then aggregate the scores from each of the comments on the post (max, mean, p95, p75, p50) and use each of those aggregates as features.

We found that a combination of our linear regression and DeepText models made the most effective classifier. In aggregate, text-related features are given 80 percent importance, with other circumstantial features (such as the reaction counts) making up the other 20 percent for our most performant classifier. We run a mix of these classifiers in order to maximize recall.

To broaden coverage across more languages, we use a combination of multilingual and monolingual classifiers, and we are working on utilizing DeepText's cross-lingual abilities to further improve the system's performance. Some languages have more training data than others. This technology allows us to leverage languages with larger data sets to cover languages with smaller sets by associating words and phrases that have the same or similar meanings in multiple languages into similar vector representations. This enables us to leverage larger amounts of training data in certain languages into high-quality classification in other languages, with little or no training data.

To be sure, integrating this technology into our suicide prevention tools does not eliminate people from the process, and it never will — instead, it amplifies their impact. Friends and family are still at the center of this technology, since posts reported by friends and family and the comments associated with them are the strongest signal we have to indicate that someone is at risk. Then after a post is detected, it is reviewed by a Community Operations team member, who then decides whether the person who posted may be in need of support resources. At this point, the classifier is not 100 percent effective in identifying every post truly expressing suicidal ideas but we are improving our efforts to get our community help in real time by using artificial intelligence to support our work.

We have more work to do, but with the help of our partners and friends and family on Facebook, we are hopeful we can support more people over time.

TAGS: USER EXPERIENCE



◂ **Prev**
F8 2018 registration now open

Next ▸
Global connectivity: Working together to bring more people online



Read More in ML Applications

View All ▸



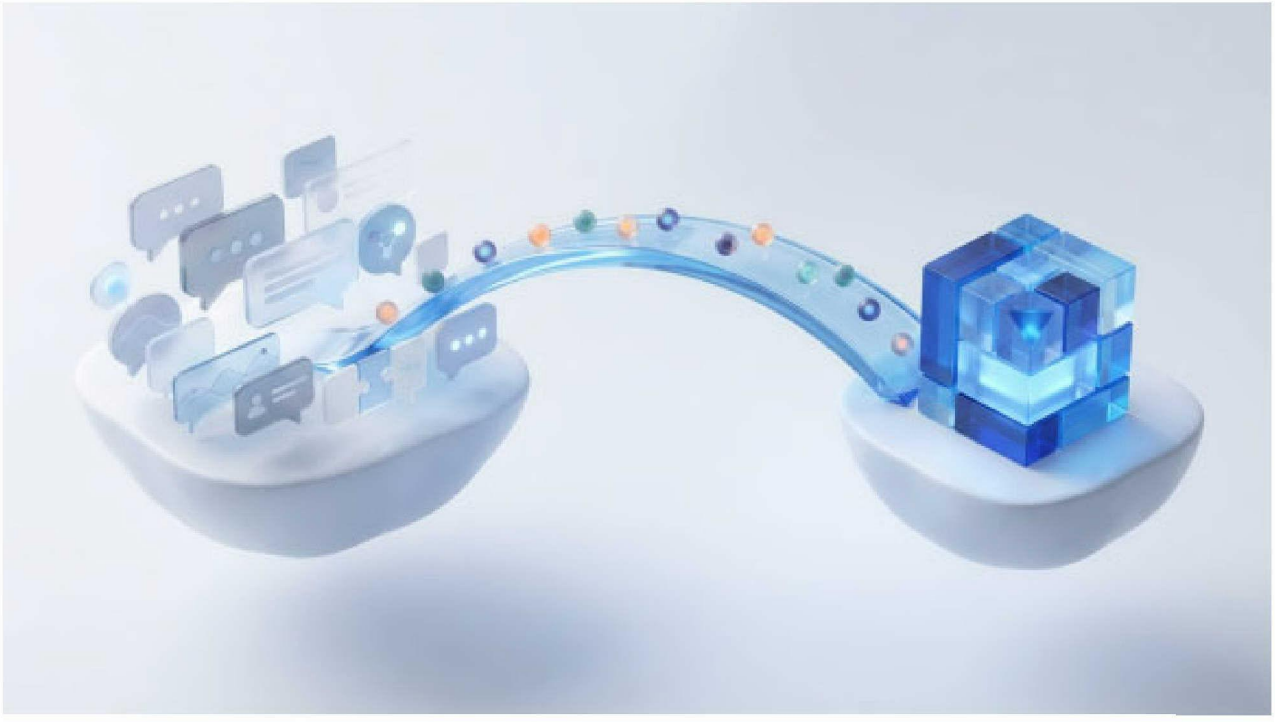
MAY 26, 2026

SilverTorch: Index as Model — A New Retrieval Paradigm for Recommendation Systems



MAY 13, 2026

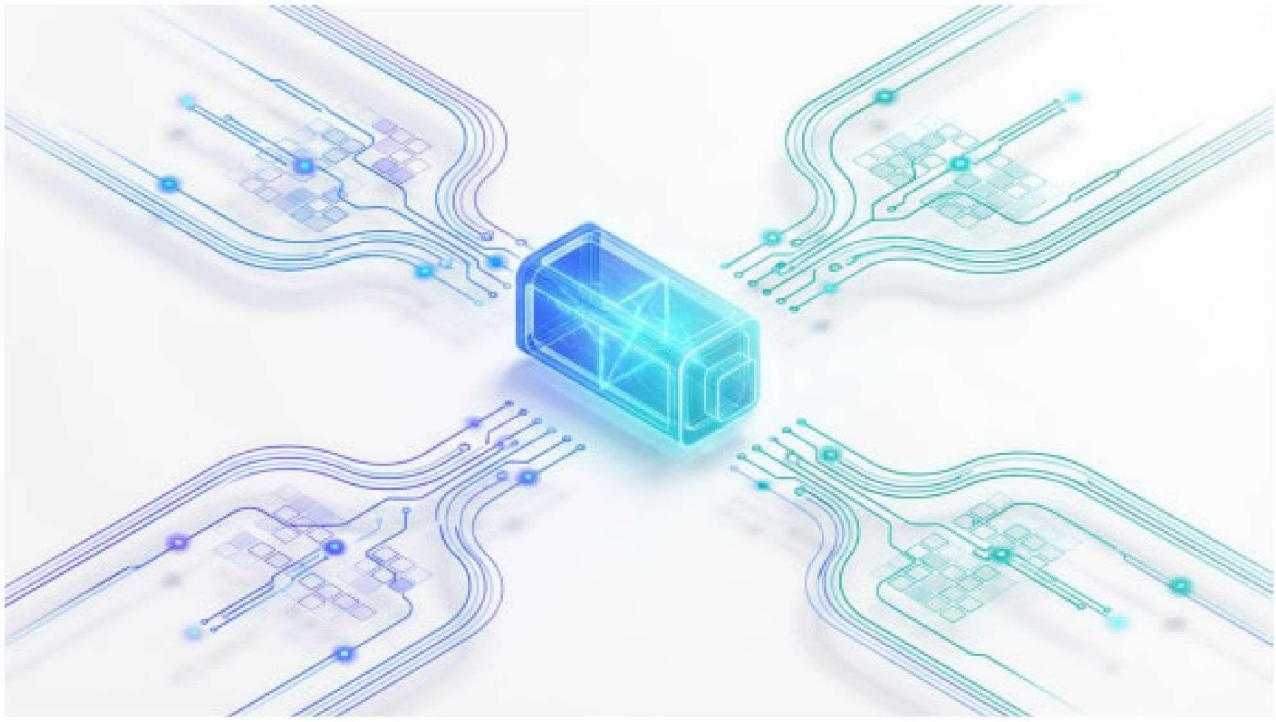
Reel Friends: Building Social Discovery that Scales to Billions



APR 21, 2026

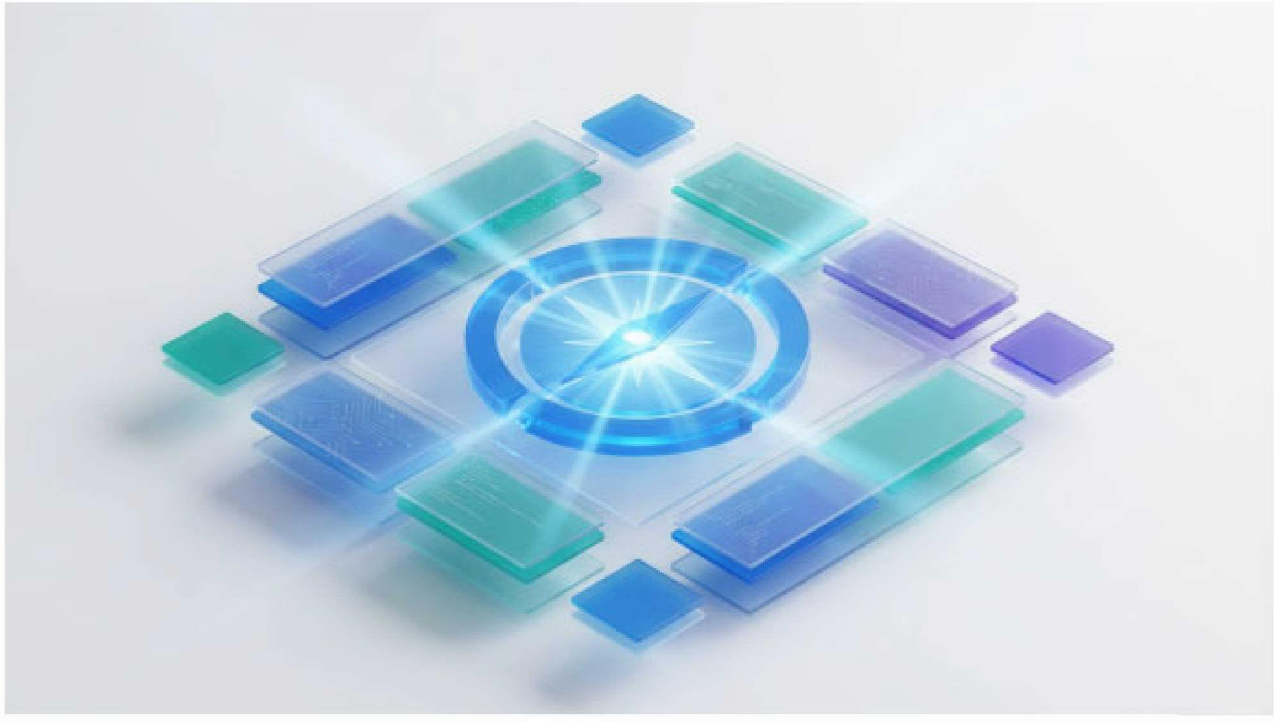
Modernizing the Facebook Groups Search to Unlock the Power of Community Knowledge

https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/



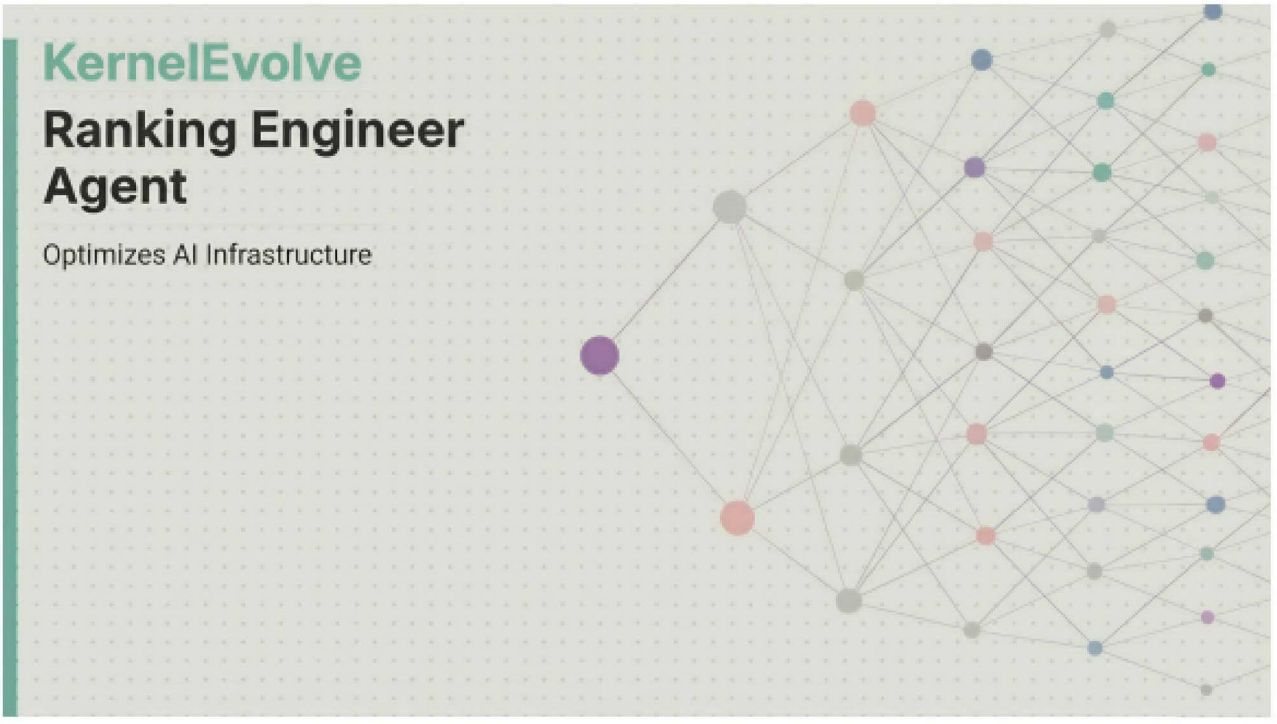
APR 16, 2026

Capacity Efficiency at Meta: How Unified AI Agents Optimize Performance at Hyperscale



APR 6, 2026

How Meta Used AI to Map Tribal Knowledge in Large-Scale Data Pipelines



APR 2, 2026

KernelEvolve: How Meta's Ranking Engineer Agent Optimizes AI Infrastructure

Related Posts

Related Positions

Software Engineer, Machine Learning
SUNNYVALE, US

Software Engineer, Machine Learning
REMOTE, US

Software Engineer, Machine Learning
SAN MATEO, US

Software Engineer, Machine Learning
BELLEVUE, US

Software Engineer, Machine Learning
REDMOND, US

See All Jobs

Available Positions

Software Engineer, Machine Learning
SUNNYVALE, US
Software Engineer, Machine Learning
REMOTE, US

Technology at Meta

Engineering at Meta - X

Open Source

Meta believes in building community through open source technology. Explore our latest projects in Artificial Intelligence,

https://engineering.fb.com/2018/02/21/ml-applications/under-the-hood-suicide-prevention-tools-powered-by-ai/

Software Engineer, Machine Learning
SAN MATEO, US
Software Engineer, Machine Learning
BELLEVUE, US
Software Engineer, Machine Learning
REDMOND, US

See All Jobs

Follow

AI at Meta
Read

Meta Quest Blog
Read

Meta for Developers
Read

Meta Bug Bounty
Learn more

RSS
Subscribe

Data Infrastructure, Development Tools, Front End, Languages, Platforms, Security, Virtual Reality, and more.

ANDROID

iOS

WEB

BACKEND

HARDWARE

Learn More

Meta

Engineering at Meta is a technical news resource for engineers interested in how we solve large-scale technical challenges at Meta.

Home

Company Info

Careers

© 2026 Meta
TermsPrivacyCookiesHelp